CLAIMS

What is claimed is:

1. A method for processing a formatted computer-readable source document having one or more pages, the source document having text comprising a plurality of words, each word comprising one or more characters, each character having a character appearance defined by one or more font properties and each word having a word appearance defined by the font properties of its characters and a position of the word on a page, the method comprising:

partitioning the formatted text into one or more groups of words and assigning at least one element from a predefined set of markup language elements to each group of words, each group of words having a group appearance defined by one or more text properties;

for each of the predefined elements that is assigned to at least one group or words, deriving an element style comprising a character style, a layout style or both, the character style being derived from the font properties of the characters of the words in the groups of words to which the element is assigned, and the layout style being derived from the text properties of the groups of words to which the element is assigned; and

creating an electronic document comprising a style sheet defining each of the element styles.

2. The method of claim 1, wherein the text is partitioned into groups of words according to word positions and an element is assigned to each group of words solely on the basis of the position of the group of words on the page.

3. The method of claim 1, wherein the text is partitioned into groups of words according to the font properties of the words in the text and an element is assigned to each group of words solely on the basis of the font properties of the words in the group of words.

4. The method of claim 1, wherein the text is partitioned into groups of words according to word positions and the font properties of the words in the text and at least one element is assigned to each group of words based on the position of the group of words on the page and the font properties of the words in the group of words.

5. The method of claim 1, further comprising:

for each element assigned to at least one group of words, comparing the group appearances of all groups of words to which the element is assigned, creating one or more alternate elements if the differences among the group appearances exceed a predefined threshold, and assigning each group of words to the original element or an alternate element.

6. The method of claim 5, wherein a numeric value defining the predefined threshold is obtained from user input.

7. The method of claim 5, wherein a numeric value defining the predefined threshold is a preprogrammed numeric value.

8. The method of claim 1, wherein the set of predefined elements is the set of HyperText Markup Language elements defined in HTML 4.0.

9. The method of claim 1, wherein the set of predefined elements is the set of Extensible Markup Language elements defined in XML 1.0.

10. The method of claim 1, wherein the predefined set of elements comprises:

a header element and a paragraph element.

11. The method of claim 10, wherein the predefined set of elements further comprises:

an address element, a blockquote element, a list element, a table element and a caption element.

12. The method of claim 1, wherein the character style comprises at least one font property and an associated value.

13. The method of claim 12, wherein the font property is selected from a predefined set of font properties comprising a font family, a font style, a font weight, a font variant and a font size.

14. The method of claim 1, wherein the layout style comprises at least one text property and an associated value.

15. The method of claim 14, wherein the text property is selected from a predefined set of text properties comprising a text decoration, a text alignment, a text indentation and a text transformation as defined in CSS1.

16. The method of claim 1, wherein the source document is derived from a raster image of the text processed by an optical character recognition (OCR) process.

17. The method of claim 16, further comprising translating the raster image of the text into a HTML description of the text.

18. The method of claim 1, further comprising detecting and setting page margins for the document page.

19. The method of claim 1, wherein the style sheet is an extensible style sheet (XSL).

20. The method of claim 1, further comprising creating an electronic document comprising a markup language version of the source document.

21. A computer program product, tangibly stored on a computer-readable medium, for processing a formatted computer-readable source document having one or more pages, the source document having text comprising a plurality of words, each word comprising one or more characters, each character having a character appearance defined by one or more font properties and each word having a word appearance defined by the font properties of its

characters and a position of the word on a page, the product comprising instructions operable to cause a programmable processor to:

partition the formatted text into one or more groups of words and assign at least one element from a predefined set of markup language elements to each group of words, each group of words having a group appearance defined by one or more text properties;

for each of the predefined elements that is assigned to at least one group of words, derive an element style comprising a character style, a layout style or both, the character style being derived from the font properties of the characters of the words in the groups of words to which the element is assigned ,and the layout style being derived from the text properties of the groups of words to which the element is assigned; and

create an electronic document comprising a style sheet defining each of the element styles.

22. The product of claim 21, wherein the text is partitioned into groups of words according to word positions and an element is assigned to each group of words solely on the basis of the position of the group of words on the page.

23. The product of claim 21, wherein the text is partitioned into groups of words according to the font properties of the words in the text and an element is assigned to each group of words solely on the basis of the font properties of the words in the group of words.

24. The product of claim 21, wherein the text is partitioned into groups of words according to word positions and the font properties of the words in the text and at least one element is assigned to each group of words based on the position of the group of words on the page and the font properties of the words in the group of words.

25. The product of claim 21, further comprising:

for each element assigned to at least one group of words, comparing the group appearances of all groups of words to which the element is assigned, creating one or more alternate elements if the differences among the group appearances exceed a predefined threshold, and assigning each group of words to the original element or an alternate element.

A10
A17